Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

The Role of Somatosensory Models in Vocal Autonomous Exploration

El rol de los sistemas somatosensoriales en exploración, vocal temprana

Acevedo-Valle, J. M.[1], Angulo, C.[1], Moulin-Frier, C.[2], Trejo, K.[1]

[1] GREC; Universitat Politècnica de Catalunya; Barcelona. Spain.
juan.manuel.acevedo.valle@upc.edu; cecilio.angulo@upc.edu;
karla.andrea.trejo@upc.edu

[2] Flowers Team, Inria/ENSTA-Paristech, Bordeaux, France[1].
clement.moulinfrier@gmail.com

Technological Innovation: This work applies machine learning techniques to investigate autonomous speech intelligent systems.

Industrial Application Area: Natural Language User Interfaces, Robotics Industry, Rehabilitation Assistance.

Recibido: 27 junio 2016.
Aceptado: 15 octubre 2016.

Abstract

The present work focuses on two main objectives. Firstly, it highlights the relevance of studying the early stages of language development using machines as an approach to contribute to the future of speech recognizers and synthesizers, user interfaces, active learning techniques, and to the field of robotics and artificial intelligence in general. Secondly, this work introduces some results on the study of the role of somatosensory models in vocal autonomous exploration. In previous works, the roles of intrinsic motivations and motor constraints in early vocal development were studied showing that active learning techniques can be used by artificial agents endowed with a simulated vocal tract to autonomously learn how to produce intended sounds through the use of probabilistic models. This work studies the effects of modifying the somatosensory model, which is used to map motor commands to undesired articulatory

[1] C. Moulin-Frier is now at SPECS, Universitat Pompeu-Fabra, Barcelona, Spain.

configurations, over the intrinsically motivated active learning process. The somatosensory system is modeled as a Gaussian Mixture Model. Herein, some simulations were run varying the structure of the model in order to analyze differences in the results. The effects on the explored sensorimotor regions and the amount of undesired vocal configurations are studied. The simulations presented in this work show that the structure of the current somatosensory model is relevant to the learning process. However, it can be also concluded that in order to reliably characterize the effects of modifying the somatosensory model further simulations must be performed and clear measures for performance should be considered.

Key Words: Autonomous Vocal Exploration, Speech Technologies, Active Learning.

Resumen

El trabajo presentado persigue dos objetivos principales: el primero de ellos es mostrar la necesidad de estudiar las etapas tempranas del desarrollo del lenguaje utilizando máquinas. Estos estudios contribuirán en el desarrollo futuro de sintetizadores y reconocedores de voz, interfaces de usuario e indirectamente al estudio de la inteligencia artificial; el segundo objetivo es presentar nuevos resultados en el estudio sobre el rol de los sistemas somatosensores en la exploración vocal temprana. En trabajos preliminares fueron estudiados los roles de las motivaciones intrínsecas y las restricciones motoras en el desarrollo vocal temprano. De estos estudios se concluyó que las técnicas de aprendizaje automático activo pueden ser utilizadas en conjunto con agentes artificiales dotados con un tracto vocal simulado para aprender autónomamente cómo producir sonidos específicos. En el presente trabajo se estudian los efectos del cambio de los parámetros que definen el modelo probabilístico del sistema somatosensorial, el cual mapea configuraciones motoras con configuraciones articulares indeseadas sobre el proceso de aprendizaje. El sistema somatosensorial es modelado utilizando *"Gaussian Mixture Models"*. A través del resultado de una serie de simulaciones donde se modifica la estructura del modelo antes mencionado, se demuestra que la estructura del modelo somatosensorial es relevante para el proceso de aprendizaje. Sin embargo, los resultados también indican que para realizar una mejor caracterización de los efectos de la modificación del modelo somatosensorial deben llevarse a cabo más simulaciones, así como tomar en consideración nuevas medidas de calidad del aprendizaje.

Palabras Clave: Exploración Vocal Autónoma, Voz Artificial, Aprendizaje Activo.

1. Introduction

During the past decade, the field of robotics has been exponentially growing [1]. Some of the challenges that robots and robot builders will face in the future is elaborated in [1]. Furthermore, it is stated that a revolution in the field of robotics was and is still occurring, and it will lead civilization to an age where robots will be in every home, every office and every work place. In short, robots will persist in everyday human life. Currently the best example is the manufacturing industry, where robots have taken a core role in satisfying the needs of humans.

Furthermore, the field of robotics is overcoming the previously persistent lack of standardization. Tools like the Robotic Operating System (ROS) middleware and more generic software tools like Open Source Computer Vision (OpenCV) have emerged as powerful tools for the development of robots. Also, important

initiatives such as the Google Summer of Code, the DARPA project, and the RoboCup competitions are fueling the evolution of tools applicable to robotics.

In spite of this optimistic view on the future of robotics, robots have still not pervaded our daily life. A number of scientific issues are yet to be solved for robots to be able to efficiently behave in open and uncertain environments. Modern approaches like Developmental Robotics aim at solving some of these issues by understanding and modeling key cognitive processes in robots.

Among the most challenging fields associated to robotics, one could find computer vision, navigation, motion control, and human robot interaction. To build robots capable of interacting efficiently and safely with humans, engineers must focus on building complex, interactive machines endowed with human-like interaction mechanisms. One of the most important interactive mechanisms is speech. However, providing a machine with the ability to use natural language through speech is not an easy task. In addition, endowing robots and machines in general, with shared intentionality features is an equally challenging task [2]. Finally, in order to understand the challenges faced to endow a machine with advanced speech human interaction tools, the reader is referred to [3], [4], and [5].

The present work is a result of an on-going research project related to [6] and its extended version (pending of publication). Therein lies a developmental robotics approach which is applied to study early-vocal development using a state-of-the-art vocal tract-ear simulated model and an artificial cognitive architecture. developmental robotics aims at understanding the key developmental processes which allow the progressive acquisition of an efficient interaction of an embodied agent with the environment, including social peers. The approach involves modeling concepts from developmental sciences into embodied robots, both to validate and confirm existing hypotheses and to propose new original ones [7].

The embodiment paradigm, also represented by the quote “understanding by building”, states that the behavior of biological or artificial agents is not only the result of a system control structure itself, but it is also affected by the ecological niche, the morphology and material properties of the agent [8]. Previously in [6], we presented some results under the hypothesis that early-language acquisition can be studied as a result of embodied cognition and developmental mechanisms produced by human evolution. This kind of work is an attempt to bridge two main ideas that could contribute to generate human-like speech technologies. On the one hand, the idea is that human language acquisition is constrained by motor, perceptual, social, and learning abilities [3]. On the other hand, it is considered that intrinsically motivated exploration algorithms are crucial components of the cognitive architectures that allows humans to acquire language [9].

The main contribution of this work refers to the analysis of simulations using the initial algorithm presented in [6] for the case of intrinsic motivated vocal exploration. That work introduced a somatosensory model which endows the embodied agent with constraint awareness. In mammals and other vertebrates, the somatosensory systems consist of nerve cells that respond to changes into the surface or internal state of the body; the collected information is sent to the brain and processed to generate a self-body image. In this work, the somatosensory model, based on the vocal-

tract surface, is used to predict violations to motor constraints. When a violation is expected, then the motor command is not executed and the searching area is moved. Moreover, the parameters of the somatosensory model are varied to modify its structure and to analyze the effects of this new parametrization on the exploration progress.

The remainder of this paper is organized as follows. The second section presents a brief description of artificial intelligent speech technologies. The concept of artificial early-vocal development and intrinsic motivations is introduced in the third section. Whereas the fourth section is aimed at explaining the algorithm presented in [6] for autonomous active exploration. Finally, the fifth and sixth sections are devoted to results and conclusions respectively.

2. Intelligent Artificial Speech Technologies

In [16], Human-Robot Interaction (HRI) and Human-Machine Interaction (HCI) systems are approached as two different problems. It is also mentioned how complex mechanical bodies provide physical features that gives HRI an advantage over HCI. In the latter, the dialogue is the most prominent matter of communication. Thus, in the long term, this project aims at addressing multi-modal robot-interaction where speech would be reinforced with other levels of HRI and cooperation. The reader is referred to [17] for more details on the state-of-the-art speech technologies.

Speech recognition technologies are not new players in the artificial intelligence field. They have long been used for biometric purposes, dictation systems, telephonic spying, and human-robot interaction systems, to name a few examples. However, these technologies face many challenges that are still unresolved today.

The knowledge on human speech processing is still very limited. In [10], twenty questions considered to be important to achieve a greater understanding of the nature of speech mechanisms and speech pattern processing were formulated. Few years later the answer to those question remains unclear [11]. The aim of this work and its overarching project is contributing towards answering the following questions: How important is the communicative nature of speech? Speech technology or speech science? How much effort does speech need? What is a good architecture for speech processing? How important are physiological mechanisms? What are the mechanisms for learning? What is speech good for? How good is speech? It is expected to answer these questions through the design of learning architectures based on the evidences mentioned in [3], [4], [5], and [12].

Looking at the recent research on artificial speech technologies is enough to realize that they are still a heavily contested topic. In [13], a natural language processing technology using Microsoft Kinect is used to produce an interactive system with human-oriented operation intuitions. In [14], the problem of speaker localization in noisy environments is approached under the hypothesis that the actual challenge is to find human-like speech sources over predominant sound sources. In [15], a web Multi-Modal User Interface (MMUI) for elderly users is presented, however the speech interaction interface is limited to a pre-specified grammar, similar to many other speech interface applications. The future of speech interaction systems must be driven towards more natural and constraint-relaxed grammar and speech perception.

3. Artificial Early-vocal Development and Intrinsic Motivations

As mentioned in [18], human speech production and perception are among some of the most complex processes occurring in living beings. Speech production requires coordinated control of many degrees of freedom for the respiratory, laryngeal and supraglottal articulatory systems to produce a linguistic message that can be understood by another human endowed with specialized perception mechanisms. Action and perception speech mechanisms are developed in the early stages of human life.

The structure of vocal development process in these early stages is the same for all typical developing individuals [3]. The infant first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds and finally automatically discovers and focuses on babbling with articulated proto-syllables. Previous studies attempting to reproduce the developmental processes of language emergence assumed the existence of structural developmental stages. However, they bridged those stages using hard-coding for experimentation [19, 20].

In [9], an attempt to explain the nature of developmental stages as a result of self-organization was presented. The work implemented a novel artificial cognitive architecture to study emergence of early-vocal development stages in machines. Their results suggest that intrinsic motivation allows infants to progressively learn to control their vocal tract. The reader is referred to [21] and [22] for further details on the role of intrinsic motivations in robotics.

4. A Cognitive Approach to Study Somatosensory Systems for Vocal Development

In [6], vocal exploration is approached by applying intrinsically motivated algorithms. Different from [9], [6] included the autonomous learning of motor constraints for which an artificial somatosensory system was introduced in the learning process.

The exploration algorithm for early-vocal development in [6] considers four elements:

- The sensorimotor and somatosensory systems are represented by the DIVA model [19] and correspond to the physical properties of the embodied agent.
- The sensorimotor model is a map from motor commands to auditory results.
- The interest model allows the agent to actively choose auditory goals which are likely to improve the quality of its sensorimotor model.
- The somatosensory model is the self-awareness mechanism for mapping physical constraints to motor commands.

4.1 Sensorimotor and Somatosensory Systems

This work reproduces the experimental setup proposed in [9] and extended in [6]; for further details on the implementation the reader is referred to those papers. There, the vocal-tract of the DIVA model is used as a sensorimotor system, the vocal tract shape is determined by the position of ten articulators and three voicing parameters. As in [9], only seven articulators and two voicing parameters are considered for experimental purposes.

Articulators and voicing parameters dynamics are modeled as over-damped second order systems: $\dot{x} + 2\zeta\omega + \omega(x - m) = 0$, with $\zeta = 1.0$, and $\omega = 2\,\pi/0.8$ representing the damping factor and the natural frequency, respectively. The fixed duration of each vocal experiment, or vocalization, is 0.8 seconds, whereas m and

x represent the motor command and the current articulator position, respectively.

Each vocalization is divided in two parts with a specific motor command, the first from 0 to 0.25 seconds, and the second from 0.25 to 0.8 seconds. Thus, linking both time windows the result is an 18-dimensional motor command vector, considering there are 7 articulators and 2 voicing parameters.

Human speech can be described by its formant frequencies. The first two formant frequencies, F_1 and F_2, are considered here along with an intonation parameter I(considered 1 when phonation occurs and 0 otherwise) to characterize the auditory result of an experimental vocalization.

During each vocalization the auditory result is observed along two time windows: the first from 0.25 to 0.4 seconds, and the second from 0.65 to 0.8 seconds. The value of each auditory output is averaged for each time window of which the result is a 6-dimensional auditory output signal (2 formants and the intonation, hence 3 values, for each of the 2 time windows).

For the somatosensory system, the minimal value of the area function $min(a_f)$ is considered, where a_f is a numerical vector that describes the shape of the vocal tract. When a value of a_f is zero, it means that the vocal tract is blocked and thus no sound is produced. On the other hand, when it contains a negative value it means that some tissues are overlapped which does not make physical sense. Viewed in another way, it might be interpreted as an excess of pressure between the tongue and the palate, the tongue being bitten, or any other uncomfortable or painful event. Hence, the average value of $min(a_f)$ is computed for both time perception windows to generate a proprioceptive feedback signal p: if the average of $min(a_f)$ is lower than a threshold for any perception window, then $p = 1$ which means an undesired configuration has occurred and $p = 0$ otherwise.

4.2 Sensorimotor Model

Gaussian Mixture Models (GMM) are linear combinations of multivariate Gaussian distributions that represent clusters of data. In this work, systems are modeled as GMMs which are trained using the algorithms based on the open source tools associated to [23]. Training is done using data collected from vocalization experiments with the DIVA system.

To build the models an n -dimensional input command space $X \in R^n$ is mapped to an m-dimensional output space $Y \in R^m$, considering the function $y = f(x) + \varepsilon$, where $y \in Y$, $x \in X$ and ε is white noise. When a batch of couples (x, y) is available, the modified Expectation-Maximization (EM) algorithms are used to obtain a GMM distribution described by the parameters $\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}$, where π_j, μ_j and Σ_j are the prior probability, the distribution centroid and the covariance matrix of the j-th Gaussian, respectively, for $j = 1,2, \dots, K$, being K the number of Gaussian distributions in the mixture. The training procedure includes a learning rate parameter α. Bayesian inference is used to compute the conditional probability distribution $P(X|y)$ in the input space X given a desired output y [9, 23] and select the value $x^* \in X$ that maximizes $P(X|y)$.

For the sensorimotor model, an 18-dimensional motor command space M is defined for the vocal tract articulatory configuration. Also, a 6-dimensional auditory output space S is defined. The observations $s \in S$ are generated according to $s = f(m) + \varepsilon$, where $\varepsilon = N(0,0.01)$ is Gaussian noise. A GMM G_{SM} with K_{SM} components representing the inverse model

$m = f^{-1}(s_g)$ must be found, where s_g is an auditory goal.

To compute G_{SM}, the equivalences $X = M$ and $Y = S$ are assumed. It allows to compute the inverse model $P(M|s_g)$ using Bayesian inference. At the beginning of the experiment, m is selected randomly to initialize the inverse sensorimotor model $P(M|S)$ around a specific region of the sensorimotor space. After the initialization stage, the agent starts to select new auditory goals s_g according to the interest model explained below and computing motor commands according to $m \approx f^{-1}(s_g) \approx P(M|s_g)$. G_{SM} is trained every N_{SM} vocalizations.

4.3 Interest Model

To build the interest model, two different auditory spaces are considered: the space of actual auditory productions and the space of auditory goals. The agent's competence is defined as $c = e^{|s_g - s|}$ [9], where s is the actual auditory production after executing a motor command $m \approx P(M|s_g)$. The interest model G_{IM} for auditory goals with K_{IM} Gaussian components represents the evolution in time of the competence over the auditory goal space S. It is used to compute a probability distribution $G_{IM}(S)$, which represents how likely an auditory goal sample is to contribute positively to the agent's competence to produce intended sounds.

To construct the interest model, the auditory goal space is augmented with two extra dimensions: the competence $c \in C$ and a time tag $t \in T$. The model G_{IM} which is a GMM with K_{IM} components will be computed using a batch of N_{IM} samples of the augmented goal space. To initialize this model some auditory results from the initialization of G_{SM} are selected as the first auditory goals s_g.

The Gaussian components in G_I that contain goals which are likely to increase the competence are considered to build the probability distribution $G_I(S)$ over the auditory space. In $G_{IM}(S)$, Gaussian distributions are weighted according to the magnitude of the covariance between time and competence. Thus, $G_{IM}(S)$ prioritizes goals in regions where competence is expected to increase. Finally, a sample s_g might be drawn from $G_{IM}(S)$. Model training is performed every time the agent has performed n_{IM} experiments.

4.4 Somatosensory Model

Different from previous architectures for sensorimotor exploration, we include constraint awareness through the usage of somatosensory information. To the best of our knowledge, there are no available cognitive architectures in the literature for the same purposes that explicitly consider motor constraints. Somatosensory information is collected from the changes in shape of the vocal-tract.

Considering the 18-dimensional motor command space M, with $m \in M$ and a proprioceptive output space $P \in [0,1]$, with $p \in P$, where p is the output of the somatosensory system. A function $g(\cdot)$ is assumed to exist such that $p = g(m)$ and the agent can observe p for each vocal experiment. Thus, a GMM model G_{SS} with K_{SS} Gaussian components can be computed with $X = M$ and $Y = P$, that allows to compute the probability distribution $G_{SS}(P|m)$ applying Bayesian inference and predict how likely is that m leads to an undesired collision, i.e. the proprioceptive feedback p is expected to be 1.

The somatosensory model is initialized together with the sensorimotor model. When an auditory goal s_g has been selected, m is computed with $P(M|s_g)$. Next, $G_{SS}(P|m)$ is

used to predict the value of p_{tmp}, if it is close to 1 then s_g is rejected, otherwise s_g and m are accepted. If s_g is rejected, then $G_{IM}(S)$ is recomputed neglecting the Gaussian distribution that generated s_g. The new $G_{IM}(S)$ is used to select a new goal s_g and the process is repeated until s_g is accepted. Every time G_{SM} is trained, G_{SS} is also trained.

4.5 Exploration Algorithm

The self-exploration algorithm is the one presented in [6]. Compared to the algorithm in [9], Algorithm 1 is augmented with lines 5-12 and line 18. Algorithm 1 starts when no experience exists in vocalizing after which the models G_{SM}, G_{SS} and G_{IM} are initialized as mentioned in previous sections.

Self-exploration with goal babbling and self-constraints awareness.

1: Set $\{K_{SM}, K_{IM}, K_{SS}, \alpha_{SM}, \alpha_{IM}, N_{SM}, N_{IM}$, and $n_{IM}\}$
2: Initialize G_{SM} and G_{SS}
3: Initialize G_{IM} and $i \leftarrow 0$
4: **while** true **do**
5: $p_{tmp} \leftarrow 1$
6: **while** p_{tmp} **do**
7: $s_{g,i} \leftarrow G_{IM}(S)$
8: $m_i \leftarrow G_{SM}(M|s_{g,i})$
9: $p_{tmp} \leftarrow G_{SS}(P|m_i)$
10: **if** p_{tmp} **then**
11: $update(G_{IM}(S))$
12: **end if**
13: **end while**
14: $s_i \leftarrow f(m_i) + \sigma_{noise}$ and $p_i \leftarrow g(m_i)$
15: $c_i \leftarrow e^{-|s_{g,i} - s_i|}$
16: $i \leftarrow i + 1$
17: **if** $i \bmod N_{SM} = 0$ **then**
18: $train\left(G_{SM}, m_{(i-N_{SM}+1:i)}, s_{(i-N_{SM}+1:i)}\right)$
19: $train\left(G_{SS}, p_{(i-N_{SM}+1:i)}, s_{(i-N_{SM}+1:i)}\right)$
20: **end if**
21: **if** $i \bmod n_{IM} = 0$ **then**
22: $train\left(G_{IM}, s_{g,(i-N_{IM}+1:i)}, c_{(i-N_{IM}+1:i)}\right)$
23: **end if**
24: **end while**

Alg. 1 Exploration Algorithm from **[6].**

When the system has been initialized, the agent draws a goal s_g for the next experiment from $G_{IM}(S)$ and the motor command m is obtained. Then, $G_{SS}(P|m)$ is used to compute p_{tmp} which indicates if the selected m may lead to an undesired collision. If p_{tmp} is close to 1, then the goal is rejected and the probabilistic distribution $G_{IM}(S)$ is updated. Otherwise, if p_{tmp} is close to zero, both s_g and m are accepted. Next, the motor command is executed with the vocal tract and the agent observes s and p, then it evaluates c. It also checks if we are at the end of a learning episode, in that case G_{SM}, G_{SS}, and G_{IM} are trained.

5. Results

Twelve different simulations were run using Algorithm 1. Three different sets of initial conditions with 42 vocalization samples were used and for each set four different simulations with one million vocalizations were executed varying the number of Gaussians $K_{SS} \in \{14, 28, 35, 42\}$ for the somatosensory model G_{SS}. Other important parameters were kept as in [6] and [9] which are $K_{SM} = 28$, $N_{SM} = 400$, $K_I = 12$, $N_I = 4800$, $n_I = 12$, and the sampling time for the DIVA model is $t_s = 0.01$ seconds. The learning rate parameter α for G_{SM} begins from 0.1 and decreases logarithmically to 0.05 after one million vocalizations whereas, for G_{SS}, $\alpha = 0.05$ is preserved through the whole simulation.

The results are evaluated using Table 1 and Figures 1-3. Table 1 shows the mean competence and the total percentage of undesired contacts through the whole simulation. Whereas Figures 1-3 represent auditory output space projections of explored regions during the simulations. One figure is depicted per each initial set of vocalizations and the explored regions using different values for K_{SS} are represented in different colors. They are stacked according to the K_{SS} value, the greatest K_{SS} appears in the front.

Table 1. Competence and contacts.

Initial Set	K_{SS}	$mean(c)$	% Contacts
1	14	0.9096	3.26%
1	28	0.92358	1.26%
1	35	0.90955	2.62%
1	42	0.92162	2.08%
2	14	0.8993	2.73%
2	28	0.89877	3.45%
2	35	0.90166	4.10%
2	42	0.90434	1.75%
3	14	0.87865	3.09%
3	28	0.86252	4.30%
3	35	0.86036	2.11%
3	42	0.89779	4.70%

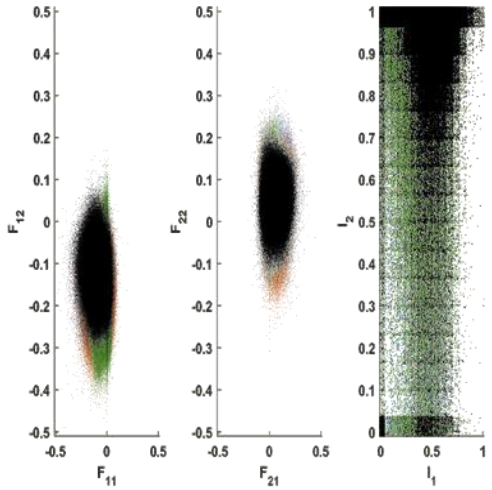

Fig. 1. Explored auditory space region using the first set of initial conditions. Blue $K_{SS} = 14$, orange $K_{SS} = 28$, green $K_{SS} = 35$, black $K_{SS} = 42$.

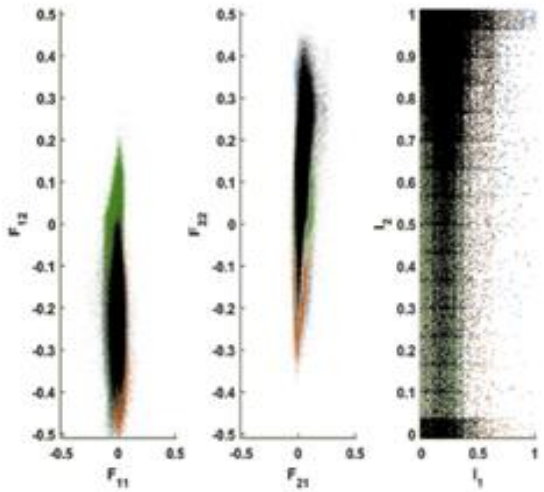

Fig. 2. Explored auditory space region using the second set of initial conditions. Blue $K_{SS} = 14$, orange $K_{SS} = 28$, green $K_{SS} = 35$, black $K_{SS} = 42$.

In Figures 1-3, $F_{11}F_{12}$ is the projection over the plane represented by the first frequency formants in the auditory perception windows, whereas $F_{21}F_{22}$ projection corresponds to the second frequency formant. Finally, I_1I_2 is the intonation projection of the auditory output. As I is a binary value (1 when vocalization is phonatory and 0 otherwise), when it is averaged within a perception window of 0.15 seconds using a sampling time of 0.01 then it only has fifteen possible values. If we plot I_1I_2 directly from the average values of I, then it would be impossible to read the results. Therefore, random small values were added to the average values of I in order to make the results more legible.

Looking at Table 1, it is possible to observe that the average competence achieved at the end of the simulation was directly proportional to the number of Gaussian distributions in G_{SS}. However, the slight improvement in competence is shadowed. First, by an unclear reduction of the total percentage of contacts at the end of the simulation. Secondly, by the increment in computational cost due to the complexity of training a GMM with more components. It can be also observed in Table 1 the simulations considered the best ones for each initial setup, as highlighted. There is no clear relationship between K_{SS} and results improvement regarding the final percentage of contacts. On the other hand, looking at Figures 1-3, it is clear that there is no obvious relationship between the volume of the auditory space explored with respect to the number of Gaussian distributions used for the somatosensory model G_{SS}. However, it is observed that the explored region is bigger when the number of Gaussian distributions considered is $K_{SS} = 28, 35$.

6. Conclusions

Previously, [6] studied the benefits of including the learning of constraints during early-vocal intrinsically motivated exploration. It demonstrated the advantages of introducing the somatosensory system and model in the exploration process. Constraint awareness provided by the somatosensory model is a mechanism to decrease redundancy in the explored sensorimotor region and avoid exploring useless regions.

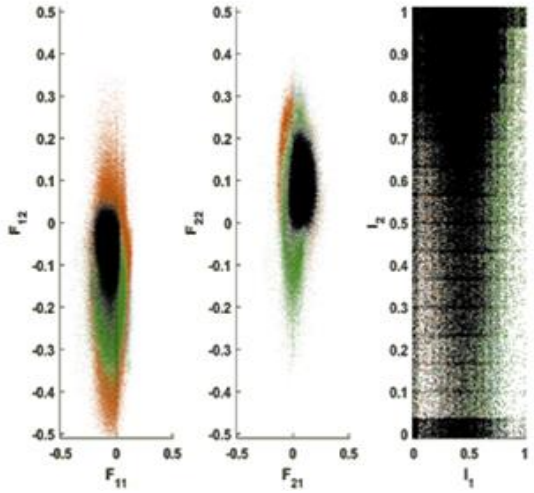

Fig. 3. Explored auditory space region using the third set of initial conditions. Blue $K_{SS} = 14$, orange $K_{SS} = 28$, green $K_{SS} = 35$, black $K_{SS} = 42$.

This work presented new results to analyze the effects on the exploration progress when the structure of the somatosensory model is modified.

Analyzing the simulation results presented in this work, it can be concluded that increasing the number of Gaussian distributions into the model of the somatosensory system is not a guarantee of better results. On the other hand, when the number of Gaussian distributions is high (e.g. 42), the computational cost for model training is significantly increased, the percentage of contacts through the simulation augmented, and the explored region shrunk. The only increased performance indicator was the competence, but the increment was not significant enough to justify the counter-effects over the computational costs.

From this work it is possible to conclude that further simulations varying parameters of the exploration algorithm must be performed to understand it in greater depth. Moreover, as mentioned in the previous work, the somatosensory model and the auditory perception system must be improved upon to be more similar to that of human beings. The need of clear measures to evaluate the quality of results obtained from an artificial agent's life in a certain simulation is left open.

7. Aknowledgement

This research is partially supported by the PATRICIA (TIN2012 - 38416 - C03 - 01) Research Project, funded by the Spanish Ministry of Economy and Competitiveness. J.M. Acevedo-Valle and K. Trejo acknowledge the received financial support from CONACYT grant 216832 and 477498, respectively.

8. Bibliography

[1] Gates, B., "A robot in every home", *Scientific American*, vol. 296, no. 1, pp. 58-65, 2007.

[2] Thompson, J. J.; Sameen, N.; Bibok, M. B.; and Racine, T. P., "Agnosticism gone awry: Why developmental robotics must commit to an understanding of embodiment and shared intentionality", *New Ideas in Psychology*, vol. 31, no. 3, pp. 184-193, 2013.

[3] Kuhl, P., "Early language acquisition: cracking the speech code", *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831-843, 2004.

[4] Galantucci, B.; Fowler, C. A.; and Turvey, M. T., "The motor theory of speech perception reviewed", Psychonomic bulletin & review, vol. 13, no. 3, pp. 361-377, 2006.

[5] Schwartz, J. L.; Boë, L. J.; Vallée, N. and Abry C., "The dispersion-focalization theory of vowel systems", Journal of phonetics, vol. 25, no. 3, pp. 255-286, 1997.

[6] Acevedo-Valle, J. M.; Angulo, C.; Agell, N. and Moulin-Frier, C., "Proprioceptive Feedback and Intrinsic Motivations in Early-vocal Development" in Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence, volume 277 of Frontiers in Artificial Intelligence and Applications, pp. 9-18, October 2015.

[7] Asada, *M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M. and Yoshida, C., "Cognitive developmental robotics: a survey", in Transactions on Autonomous Mental Development, IEEE, vol. 1, no. 1, pp. 12-34, 2009.*

[8] Pfeifer, R.; Lungarella, M. and Iida, F., "Self-organization, embodiment, and biologically inspired robotics", *Science*, vol. 318, no. 5853, pp. 1088-1093, 2007.

[9] Moulin-Frier, C.; Nguyen, S.M. and Oudeyer, P. Y., "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation", *Frontiers in psychology*, vol. 4, 2013.

[10] Moore, R. K., "Twenty things we still don't know about speech" in *Proceedings of CRIM/FORWISS Workshop on 'Progress and Prospects of Speech Research and Technology*, 1994.

[11] Anusuya, M. A. and Katti, S. K., "Speech recognition by machine, A review", *International Journal of Computer Science and Information Security*, vol. 6, no. 3, 2010.

[12] Moulin-Frier, C.; Diard, J.; Schwartz, J. L. and Bessière, P, "Communicating about Objects using Sensory–Motor Operations: A Bayesian modeling framework for studying speech communication and the emergence of phonological systems", *Journal of Phonetics*, vol. 53, pp. 5-41, November 2015.

[13] Seongha., P; Yongho, K.; Matson, E.T.; Hyeonae, J.; Changwha, L. and Wooram, P., "An intuitive interaction system for fire safety using a speech recognition technology" in *Proceedings of the 6th International Conference on Automation, Robotics and Applications*, pp. 388-392, February 2015.

[14] Hyeontaek, L.; In-Chul, Y.; Youngkyu, C. and Dongsuk, Y., "Speaker localization in noisy environments using steered response voice power" in *IEEE Transactions on Consumer Electronics*, vol. 61, no.1, pp.112-118, February 2015.

[15] Di Nuovo, A.; Broz, F.; Belpaeme, T.; Cangelosi, A.; Cavallo, F.; Esposito, R. and Dario, P., "A web based Multi-Modal Interface for elderly users of the Robot-Era multi-robot services" in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 2186-2191, October 2014.

[16] Rigoll, G., "Multimodal Human-Robot Interaction from the Perspective of a Speech Scientist", *Speech and Computer*, pp. 3-10, 2015.

[17] *Proceedings of the 17th International Conference on Speech and Computer SPECOM*, Athens, Greece, September 2015, Editors: Ronzhin, A.; Potapova, R. and Fakotakis, N.

[18] Perkell, J.; Guenther, F.; Lane, H., Matthies; M., Payan, Y.; Perrier, P.; Vick, J.; Wilhelms-Tricarico, R. and Zandipour, M., "The sensorimotor control of speech production", in *Proceedings of the First International Symposium on Measurement, Analysis and Modeling of Human Functions*, pp. 21-23, 2001.

[19] Guenther, F.H.; Ghosh, S.S. and Tourville, J.A., "Neural modeling and imaging of the cortical interactions underlying syllable production", *Brain and language*, vol. 96, no. 3, pp. 280-301, 2006.

[20] Warlaumont, A. S.; Westermann, G.; Buder, E.H. and Oller, D.K., "Prespeech motor learning in a neural network using reinforcement", *Neural Networks*, vol. 38, pp. 64-75, 2013.

[21] Oudeyer, P. Y.; Kaplan, F. and Hafner, V. V., "Intrinsic motivation systems for autonomous mental development. Evolutionary Computation", *IEEE Transactions*, vol. 11, no. 2, pp. 265-286, 2007.

[22] Gottlieb, J.; Oudeyer, P. Y.; Lopes, M. and Baranes, A., "Information-seeking, curiosity, and attention: computational and neural mechanisms", *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585-593, 2013.

[23] Calinon, S., 2009. *Robot Programming by Demonstration*. Lausanne, EPFL Press.